Sondan Eklemeli Dillerde Gövde Tabanlı Sözcük Türü İşaretleme
Stem-based PoS Tagging for Agglutinative Languages

Necva Bölücü, Burcu Can
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi, Beytepe, Ankara, Türkiye
{necva, burcucan}@cs.hacettepe.edu.tr

***Özetçe* —Sondan eklemeli dillerin temel özelliği köke getirilen eklerle yeni sözcüklerin türetilebilmesidir. Sözcüklerin bu şekilde eklerle türetilmesi seyreklik problemine neden olduğundan ötürü bu diller için sözcük türü işaretlemeyi zorlaştırmaktadır. Bu bildiride sondan eklemeli diller için gözetimsiz olarak çalışan Bayesian Saklı Markov Model temelli iki farklı sözcük türü işaretleme modeli sunulmuştur. İlk model, etiketleme için sözcükleri kullanırken, ikinci model HPS ve Morfessor FlatCat gövdeleme ve morfolojik bölümleme sistemlerinden elde edilen sözcük gövdelerini kullanmaktadır. Sonuçlar, modellerde sözcük yerine gövdelerin kullanılmasının sözcük türlerini işaretleme başarısını artırdığını göstermektedir. Önerilen model, sondan eklemeli dil olarak Türkçe ve morfolojik olarak daha zayıf bir dil olan İngilizce üzerinde çalıştırılmış ve sonuçlar kıyaslanmıştır.**

Anahtar Kelimeler—Gövdeleme, Sözcük Türü İşaretleyici, Saklı Markov Model, Sondan Eklemeli Diller

***Abstract*—Words are made up of morphemes being glued together in agglutinative languages. This makes it difficult to perform part-of-speech tagging for these languages due to sparsity. In this paper, we present two Hidden Markov Model based Bayesian PoS tagging models for agglutinative languages. Our first model is word-based and the second model is stem-based where the stems of the words are obtained from other two unsupervised stemmers: HPS stemmer and Morfessor FlatCat. The results show that stemming improves the accuracy in PoS tagging. We present the results for Turkish as an agglutinative language and English as a morphologically poor language.**

Keywords—Stemming, Part-of-Speech Tagger, Hidden Markov Model, Agglutinative languages

I. GİRİŞ

Sözcük türü işaretleme, sözcüklerin anlamı ve sözdizimine bağlı olarak sözcüklere karşılık gelen sözdizimsel kategorilerin tespit edilmesidir. Sözcük türü işaretleme; bilgi çıkarımı, konuşma tanıma, makine çevirisi gibi birçok doğal dil işleme uygulaması için en gerekli işlemlerden biridir. Bu işlem, özellikle sözcük anlamının belirsiz olmasından dolayı zorlu olabilmektedir. Örneğin *yüz* sözcüğü cümlede bulunduğu yere göre eylem ya da isim olabilir. Sözcüğün belirli bir bağlamdaki anlamının belirlenebilmesi için öncelikle sözcük türünün belirlenmesi gerekmektedir.

Türkçe, Fince gibi sondan eklemeli diller için sözcük türü işaretleme, bu dillerin yapısından dolayı diğer dillere göre daha zordur. Bu dillerde, yeni sözcük türetmek sözcüğe eklerin eklenmesi ile gerçekleştiğinden dolayı sonsuz sayıda sözcük türetilebilmektedir [9]. Sözcüğe eklenen her ek ile yeni bir sözcük formu oluşmakta, bu da seyrekliğe neden olmaktadır.

Sondan eklemeli dillerin morfolojik bakımdan zengin olması, veri kümesinde görülmeyen sözcüklerin fazla olmasından dolayı başarı oranının düşük olmasına neden olmaktadır. Birçok çalışma, sondan eklemeli dillerin sözcük türü işaretleme işleminde morfolojik yapıyı göz ardı ederek model geliştirmiştir.

Bu çalışmada sözcük yerine dilin morfolojik yapısına daha uygun olan sözcük gövdeleri kullanan bir model sunuyoruz. Gövdeleme, bir sözcüğün çekim eklerinden arındırılması işlemidir. Örneğin, *kitapçıları* sözcüğünün kökü, *kitapçı*'dır. Gövdeleme sırasında, *lar* ve *ı* sözcükleri çekim eki oldukları için çıkarılmış, *çı* eki ise yapım eki olduğu için bırakılmıştır. *Kitapçı*, sözlükte tek başına bulunabilen, sözcük kökünden ayrı bir anlamı olan bir sözcüktür.

Gövdeleme işlemi, doğal dil işleme uygulamalarında kullanılan en temel ön işlemlerden biridir. Birçok uygulama gövdeleme yöntemlerini kullanarak sözcük sayısını azaltmayı amaçlamaktadır. Aksi takdirde modelde yer verilen sözcük kümesinde bulunmayan sözlük-dışı (out-of-vocabulary) sözcüklerden dolayı seyreklik problemi ortaya çıkmaktadır.

Bu bildiride sunduğumuz model, sözcük sayısını azaltmak amacıyla gözetimsiz olarak çalışan başka gövde bulma yöntemleri kullanarak elde edilen sözcük gövdeleri ile sözcük türü etiketleme işlemini gerçekleştirmektedir. Bu da verideki seyrekliği azaltarak sözcük türlerinin daha doğru bir şekilde işaretlenmesini sağlamaktadır.

II. İLGİLİ ÇALIŞMALAR

Sözcük türü işaretlemede gözetimsiz yaklaşımlar yaygın olarak kullanılmaktadır. Bunlardan bir kısmı sözcük türü işaretleme işlemini bir kümeleme/sınıflandırma problemi olarak görmektedir. Bu amaçla Clark [4] gözetimsiz kümeleme, Schütze [16] boyut indirgeyerek kümeleme ve Brown [12] sınıf tabanlı kümeleme ile sözcük türü işaretleme işlemini gerçekleştirmiştir.

Sonraki çalışmalarda, Saklı Markov Model (Hidden Markov Model - HMM) tabanlı yöntemler kullanılmaya başlanmıştır. HMM kullanılarak yapılan ilk çalışmalardan biri Merialdo [12]'ya aittir. Merialdo çalışmasında trigram tabanlı Saklı Markov Model kullanmıştır.

TnT isimli çalışmasında Brants [1] işaretleyici olarak ikinci derece Markov modeli ve modelin öğrenilmesinde estimatör olarak ML (Maximum Likelihood) kullanmıştır.

Johnson [10] HMM sözcük türü işaretleme modeli için farklı estimatörleri karşılaştırmıştır. Beklenti Maksimizasyonu (Expectation Maximization - EM), Gibbs Örnekleme (Gibbs Sampling - GS) ve Değişimsel Bayes (Variational Bayes - VB) estimatörlerini incelemiş ve EM estimatörünün HMM için model parametrelerini öğrenmede yetersiz olduğunu bulmuştur. Değerlendirme olarak çalışmasında Çoktan-bire (Many-to-1), Bire bir (1-to-1) ve Bilgi Değişimi (Variation of Information - VI) metriklerini kulanmıştır.

Gao ve Johnson da [5] HMM sözcük türü işaretleyicilerde kullanılan farklı estimatörleri incelemiş ve deneylerini farklı büyüklüklerdeki veri kümeleri ve farklı sayıda saklı etiketler ile gerçekleştirmiştir. Çalışmada 6 farklı estimatör kullanılmıştır: EM, VB ve 4 farklı parametre ile GS.

Goldwater ve Griffiths'in [7] çalışmasında Bayesian tabanlı HMM modeli geliştirilmiştir. Standart HMM modelinden farklı olarak parametreler Multinomial-Dirichlet dağılımı ile modellenmiş ve modelin öğrenilmesinde GS örnekleme algoritması kullanılmıştır.

Moon, Erk ve Baldridge [13] tarafından geliştirilen HMM modelinde, sayıca çok olan ama metinde sık geçmeyen sözcükler öz sözcük ve sık geçen ama sayıca az olan sözcükler işlevsel sözcük olarak tanımlanmıştır. Model bu sözcükleri tespit ederek işaretleme işlemi gerçekleştirmektedir.

III. YÖNTEM

Bu çalışmada Goldwater ve Griffiths'in [7] çalışmasında önerilen Bayesian HMM modeli sondan eklemeli dillerin yapısına daha uygun olacak şekilde değiştirilmiştir. Yeni modelde sözcük yerine, sözcük gövdeleri kullanılarak sözcük seyrekliğinin önüne geçilmesi amaçlanmaktadır.

A. Sözcük Tabanlı Bayesian Sözcük Türü İşaretleme

Standart birinci dereceden HMM yapısına sahip olan model matematiksel olarak aşağıdaki gibi tanımlanmıştır:

$$t_i|t_{i-1} = t = t^{'}, \tau^{(t,t^{'})} \propto Mult(\tau^{(t,t^{'})}) \quad (1)$$

$$w_i|t_i = t, \omega^{(t)} \propto Mult(\omega^{(t)}) \quad (2)$$

$$\tau^{(t,t^{'})}|\alpha \propto Dirichlet(\alpha) \quad (3)$$

$$\omega^{(t)}|\beta \propto Dirichlet(\beta) \quad (4)$$

t_i ve w_i sırasıyla *i*nci etiket ve sözcük olarak belirlenmiştir. Durum-geçiş dağılımı, α hiper-parametresiyle tanımlanan $Dirichlet(\alpha)$ ile parametreleri oluşturulan $Mult(\tau^{(t,t^{'})})$ ile tanımlanmakta, emisyon dağılımı ise β hiper-parametresiyle tanımlanan $Dirichlet(\beta)$ dağılımı ile parametreleri oluşturulan $Mult(\omega^{(t)})$ ile elde edilmektedir.

Yukarıda belirtilen model ile t_i şartlı dağılımı aşağıda verildiği gibi olmaktadır:

$$P(t_i|t_{-i}, w, \alpha, \beta) = \frac{n_{(t_i,w_i)} + \beta}{n_{t_i} + W_{t_i}\beta} \cdot \frac{n_{(t_{i-1},t_i)} + \alpha}{n_{t_{i-1}} + T\alpha} \cdot \frac{n_{(t_i,t_{i+1})+I(t_{i-1}=t_i=t_{i+1})} + \alpha}{n_{t_i} + I(t_{i-1} = t_i) + T\alpha} \quad (5)$$

Modelde kullanılan T etiket sayısını, W_t t etiketi için izin verilen çıktılardaki sözcük türlerinin sayısını, $n_{(t_i,w_i)}$ sözcük-etiket çiftinin sayısını, n_{t_i} t_i ile etiketlenmiş sözcük sayısını ve $n(t_{-i}, t_i)$ bigram etiketlerinin bulunma sayısını vermektedir. $I(.)$ fonksiyonu eğer argümanları doğruysa 1, değilse 0 olarak çıktı üretmektedir.

B. Gövde Tabanlı Bayesian Sözcük Türü İşaretleme

Sondan eklemeli diller, yapısından dolayı çok fazla sözcük formu içermektedir. Örneğin, *kitap* sözcüğü *kitaplar, kitaptan, kitapta, kitapsa, kitapmış, kitaptı* gibi çok fazla sayıda farklı formda görülebilir. Ancak bu sözcük formlarının hepsi isim türüne aittir.

Sözcük formu sayısının fazla olması seyrek veri problemini ortaya çıkarmaktadır. Bu problemi ortadan kaldırmak için Goldwater ve Griffiths'in [7] çalışmasında kullanılan standart Bayesian HMM model, sözcükler yerine gövdeleme modelleriyle elde edilen sözcük gövdeleri kullanılarak güncellenmiştir.

Modelin değiştirilmiş matematiksel tanımı aşağıda verilmiştir.

$$t_i|t_{i-1} = t = t^{'}, \tau^{(t,t^{'})} \propto Mult(\tau^{(t,t^{'})}) \quad (6)$$

$$s_i|t_i = t, \omega^{(t)} \propto Mult(\omega^{(t)}) \quad (7)$$

$$\tau^{(t,t^{'})}|\alpha \propto Dirichlet(\alpha) \quad (8)$$

$$\omega^{(t)}|\beta \propto Dirichlet(\beta) \quad (9)$$

t_i ve s_i sırasıyla *i*nci etiket ve gövdeleme algoritmasından elde edilmiş sözcük gövdesine karşılık gelmektedir.

Yukarıda belirtilen model ile t_i şartlı dağılımı aşağıda verildiği gibi hesaplanabilmektedir:

$$P(t_i|t_{-i}, s, \alpha, \beta) = \frac{n_{(t_i,s_i)} + \beta}{n_{t_i} + S_{t_i}\beta} \cdot \frac{n_{(t_{i-1},t_i)} + \alpha}{n_{t_{i-1}} + T\alpha} \cdot \frac{n_{(t_i,t_{i+1})+I(t_{i-1}=t_i=t_{i+1})} + \alpha}{n_{t_i} + I(t_{i-1} = t_i) + T\alpha} \quad (10)$$

Modelde standart modelden farklı olarak $n_{(t_i,s_i)}$ gövde-etiket çiftlerinin sayısını ve S_t t etiketiyle işaretlenmiş gövde türlerinin sayısını göstermektedir.

Önerdiğimiz modelde gövdeyi bulmak için 2 farklı yöntem kullanıyoruz: HPS [3] ve Morfessor FlatCat [8].

HPS[1] (High Precision Stemmer) algoritması Brychcin ve Koponik [3] tarafından önerilmiştir. Bu algoritma, özellikle bilgi çıkarımı alanında büyük başarı sağlamıştır. Model iki aşamadan oluşmaktadır. İlk aşamada sözcüğün sözdizimsel ve anlamsal bilgilerine bakarak kümeleme temelli gövde bulma

TABLO I: HPS VE MORFESSOR-FLATCAT ILE BULUNMUŞ SÖZCÜK GÖVDELERI

HPS		Morfessor FlatCat	
Sözcük	**Gövde**	**Sözcük**	**Gövde**
kıvrandığın	kıvran	yanıyordu	yan
tutkuyu	tutku	söylemedim	söyle
anlattın	anlat	sizlere	siz
kurtulmak	kurtul	önemli	önem
duvara	duvar	tutkuyu	tutku
dayanıp	dayan	dedi	de
gecenin	gece	kurtulamayacak	kurtul

algoritması, ikinci aşamada da maximum entropy sınıflandırıcı algoritması kullanılmaktadır. Model birçok farklı dil grubuna ait diller üzerinde test edilmiştir.

Morfessor[2] [8], bir gövdeleyici olarak tasarlanmamış olup, aslında sözcüklerin morfolojik bölümlenmesini amaçlamaktadır. Morfessor FlatCat, en bilinen gözetimsiz morfolojik bölümleme sistemlerinden biri olan Morfessor varyasyonlarından birisidir ve HMM kullanmaktadır. Ön işlemden geçirilmemiş veri kümesi üzerinden morfemleri tahmin eden model, aynı zamanda işlenmiş veri kümesi ile gözetimli olarak da çalışabilmektedir. Morfessor FlatCat, uygulandığı veri kümesindeki sözcüklerin morfemlerine ait kök, sonek ve önek gibi morfem türlerini de vermektedir. Bu çalışmada güncellediğimiz Bayesian HMM modelde, Morfessor FlatCat'ten elde edilen kök bilgileri de kullanılarak deneyler gerçekleştirilmiştir. Burada bulunan ve kök kategorisine ait olan morfemler tanımını verdiğimiz gövde kategorisinden farklıdır, ancak yine de kök ve gövdenin farklı etkilerini görmek açısından bu çalışmada dahil edilmiştir. Her iki durumda da veri seyrekliği problemini azaltması beklenmektedir.

Morfessor FlatCat ve HPS modelleri uygulanarak Türkçe veri kümesi için elde edilen doğru gövdelere örnekler Tablo I'de verilmiştir.

IV. ÖĞRENME

Modelin öğrenilmesinde Gibbs [6] örnekleme algoritması kullanılmıştır. Bunun için öncelikle veri kümesinde gövde bulma sistemi kullanılıp gövdeler elde edilmiştir. Ardından, veri kümesindeki tüm sözcük gövdelerine etiketler rasgele atanmış ve sırasıyla her sözcük gövdesi için tüm etiketlerin şartlı olasılıkları Denklem 10'e göre hesaplanmıştır. Elde edilen etiketlere göre şartlı olasılık dağılımından bir etiket seçilip o gövde için atandıktan sonra, bu işlem iterasyon sayısınca her gövde için tekrarlanır.

Öğrenme için izlenen bu adımlar Algoritma 1'de verilmektedir. Algoritmada *W* veri kümesindeki sözcükleri ve *etiketS* etiket sayısını göstermektedir. $random(etiketS)$, rasgele bir etiket seçer ve $etiket(s_i)$, s_i gövdesine ait etiketi göstermektedir.

V. DENEYLER VE SONUÇLAR

Veri kümesi: Bu çalışmada oluşturduğumuz sözcük türü işaretleme modelini test etmek için 2 farklı veri kümesi kullandık:

[1]HPS: https://github.com/aalto-speech/flatcat

[2]Morfessor FlatCat: https://github.com/aalto-speech/flatcat

Algorithm 1 Gövde Tabanlı Bayesian Sözcük Türü İşaretleme Modelinin Öğrenilmesi

function STEMBAYESIN($W, etiketS, \alpha, \beta, iterasyon$)
  for w_i in W **do**
    $s_i \leftarrow stemmer(w_i)$
    $w_i \leftarrow s_i$
    $etiket(s_i) \leftarrow random(etiketS)$
  for k in $iterasyon$ **do**
    for s_i in W **do**
      $etiket(s_i)$ için $P(t_i|t_{-i}, s, \alpha, \beta)$ dağılımından yeni bir etiket seç
  $return\ W$

- Türkçe: METU-Sabancı Turkish Treebank (Oflazer [14]) kullanılmıştır. Bu veri kümesinde 53751 sözcük bulunmaktadır.
- İngilizce: WSJ Penn Treebank (Marcus [11]) veri kümesinin ilk 12k ve 24k sözcüğünden elde edilmiş veri kümeleri kullanılmıştır.

Bütün deneylerde, α= 0.001 ve β=0.1 olarak atanmıştır. Deneylerde Gibbs örnekleme algoritması için 1000 iterasyon kullanılmıştır. METU-Sabancı Turkish Treebank ve WSJ Penn Treebank veri kümelerinde 41 sözcük türü etiketi bulunmaktadır. Etiket sayısını, Petrov'un [15] çalışmasından yararlanarak hem Türkçe hem de İngilizce veri kümeleri için 12'ye düşürdük ve bütün deneylerde etiket sayısı olarak 12 kullandık.

Deneyler iki değerlendirme metriği ile değerlendirilmiştir.

- Doğruluk: Modeli değerlendirmek için öğrenme sonucu elde ettiğimiz etiketler ile gold veri kümesinde bulunan etiketleri eşleştiren Çoktan-bire (Many-to-1) eşleştirme metriği bu değerlendirme için kullanılmıştır.
- VI (Variation of Information): VI etiketleme sonucu M etiketinden G etiketine doğru ve yanlış tahmin edilen etiket sayılarını ölçen bilgi teorisi tabanlı bir metriktir.

Gövde tabanlı sözcük türü işaretleme modeli hem Bayesian HMM modeli [7] ile, hem de Brown kümeleme [2] algoritması ile kıyaslanmıştır.

Gövde tabanlı sözcük işaretleme modelinin değerlendirilebilmesi için öncelikle veri kümeleri üzerinde HPS ve Morfessor FlatCat ayrı deneyler olarak çalıştırılmış ve sözcük gövdeleri elde edilmiştir. Bulunan sözcük gövdeleri hem Bayesian HMM modelinde kullanılarak, hem de Brown Kümeleme algoritmasında kullanılarak sözcük işaretleme deneyleri gerçekleştirilmiştir.

Türkçe ve İngilizce için değerlendirme sonuçları Tablo II'de verilmiştir. Tabloda H-Bayesian HMM ve H-Brown Kümeleme, HPS ile edilen gövdelerin sırasıyla Bayesian HMM modeli ve Brown Kümeleme algoritmasında kullanılması sonucu oluşan gövde tabanlı sözcük işaretleme modellerini ifade etmektedir. M-Bayesian HMM ve M-Brown Kümeleme ise Morfessor FlatCat ile edilden gövdelerin sırasıyla Bayesian HMM modeli ve Brown Kümeleme algoritmasında kullanılması sonucu oluşan gövde tabanlı sözcük işaretleme modellerini göstermektedir. Bayesian HMM ve Brown Kümeleme ise sözcük tabanlı sözcük işaretleme modellerine karşılık gelmektedir.

TABLO II: SÖZCÜK TABANLI VE GÖVDE TABANLI SÖZCÜK TÜRÜ IŞARETLEME MODELLERINDEN ELDE EDILEN SONUÇLAR

	Metu (%)		Penn 12k (%)		Penn 24k (%)	
	M-to-1	**VI**	**M-to-1**	**VI**	**M-to-1**	**VI**
Bayesian HMM	56.61	10.62	**57.52**	7,83	56.13	7.92
Brown Kümeleme [3]	54.91	10.83	53.78	7.58	54.11	7.78
H-Bayesian HMM	57.52	10.69	45.47	7,97	55.11	7.89
H-Brown Kümeleme [3]	54.35	10.58	57.42	7.55	**64.57**	**7.71**
M-Bayesian HMM	56.13	10.82	47.13	8.09	47.86	8.29
M-Brown Kümeleme [3]	**57.61**	**8.96**	53.31	**7.29**	55.9	**7.56**

Sonuçları incelediğimizde, morfolojik açıdan zengin olan sondan eklemeli dillerden Türkçe için Çoktan-bire (M-to-1) ve VI metriklerinin en yüksek sonuçlarının sözcüklerden ziyade sözcük gövdeleri kullanılarak çalıştırılan modellerden elde edildiği görülmektedir. Tablo II'e göre Türkçe için en yüksek sonuçlar, gövde tabanlı olarak çalıştırılan Brown Kümeleme algoritmasından elde edilmiştir

İngilizce için ise, daha küçük olan veri kümesinde en yüksek sonuç Çoktan-bire metriği için sözcük tabanlı Bayesian HMM modelden elde edilirken, VI metriği için Türkçe'de olduğu gibi gövde tabanlı Brown Kümeleme algoritmasından elde edilmiştir. İngilizce için daha büyük veri kümesinde, Çoktan-bire ve VI metriklerinden en yüksek sonuçlar yine Türkçe'de olduğu gibi, gövde tabanlı Brown kümeleme algoritmasından elde edilmiştir. Bu sonucu veri kümesini arttırdığımızda veri kümesinde bulunan benzersiz sözcük sayısının artışına bağlayabiliriz.

Tablo II'ye baktığımızda Türkçe için, Morfessor FlatCat modelinin etiketleme üzerinde HPS modelinden daha başarılı olduğu, İngilizce için ise HPS modelinin daha başarılı olduğu gözlenmektedir. Bu durum, Morfessor FlatCat'in gövdelemeden ziyade morfolojik bölümleme yapmasıyla, dolayısıyla gövdeden ziyade kök bularak seyrekliği HPS modeline göre daha fazla azaltmasıyla da yorumlanabilir.

Sonuçları genel olarak değerlendirdiğimizde M-Brown Kümeleme yönteminin hem İngilizce hem de Türkçe için M-Bayesian modelden daha iyi sonuç verdiği gözlemlenmektedir.

H-Brown Kümeleme yönteminin ise İngilizce için H-Bayesian Modelden daha iyi sonuç verdiği gözlenmektedir. Türkçede H-Bayesian model daha iyi sonuç vermektedir.

Sözcük tabanlı Bayesian ve Brown Kümeleme modellerine baktığımızda, sözcük tabanlı Bayesian modelin hem İngilizce hem de Türkçe veri kümeleri için Brown Kümeleme modeline göre daha başarılı olduğu görülmektedir.

Sözcük işaretleme sonuçları değerlendirilirken, kullanılan gövdeleme modelleri HPS ve Morfessor FlatCat'in başarılarının da sözcük işaretleme başarısında etkisi olduğu göz önünde bulundurulmalıdır. Burada tamamen gözetimsiz olarak sözcük türleri bulunduğundan ötürü gözetimsiz olarak çalışan gövdeleme modelleri tercih edilmiştir. Gözetimli olarak çalışan gövdeleme modelleri kullanıldığı takdirde, sözcük işaretleme sonuçlarının da iyileşmesi beklenmektedir.

VI. SONUÇ VE GELECEK ÇALIŞMALAR

Bu çalışmada, gözetimsiz olarak sözcük türlerinin etiketlenmesi için standart Bayesian HMM modeli [7] sondan eklemeli dillerin yapısına uygun olarak güncellenmiştir. Modelde, sözcük yerine sözcük gövdeleri kullanılarak veri kümesindeki seyreklik probleminin de önüne geçilmiştir. Sonuçlar, Türkçe için ve veri kümesinin boyutunu arttırdığımızda İngilizce için, sözcük yerine modellerde sözcük gövdelerinin kullanılmasının başarıyı arttırdığını göstermektedir.

Gelecek çalışmalar için amacımız, gövde bulma modeli olarak dışarıdan alınan hazır bir model yerine eş zamanlı olarak hem gövdeleme hem de sözcük türü işaretleme yapabildiğimiz bir model gerçekleştirmektir.

[3]Brown Kümeleme: http://www.cs.berkeley.edu/ pliang/software/brown-cluster-1.2.zip

TEŞEKKÜR

Bu araştırma 115E464 proje numarasıyla Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından desteklenmiştir.

KAYNAKLAR

[1] Thorsten Brants. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231, Stroudsburg, PA, USA, 2000. ACL.

[2] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December 1992.

[3] Tomáš Brychcín and Miloslav Konopík. HPS: High precision stemmer. *Information Processing & Management*, 51(1):68 – 91, 2015.

[4] Alexander Clark. Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 91–94. ACL, 2000.

[5] Jianfeng Gao and Mark Johnson. A comparison of Bayesian estimators for unsupervised hidden markov model PoS taggers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 344–352, Stroudsburg, PA, USA, 2008. ACL.

[6] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984.

[7] Sharon Goldwater and Tom Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, 2007.

[8] Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. The 25th International Conference on Computational Linguistics, pages 1177–1185, 2014.

[9] Jorge Hankamer. Finite state morphology and left to right phonology. In *Proceedings of the West Coast Conference on Formal Linguistics*, pages 41–52, 1986.

[10] Mark Johnson. Why doesn't EM find good HMM PoS-taggers. In *EMNLP*, pages 296–305, 2007.

[11] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June 1993.

[12] Bernard Merialdo. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–171, June 1994.

[13] Taesun Moon, Katrin Erk, and Jason Baldridge. Crouching Dirichlet, Hidden Markov model: Unsupervised PoS tagging with context local tag generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 196–206, 2010.

[14] Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. *Building a Turkish Treebank*, pages 261–277. Springer, 2003.

[15] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the LREC'12*. European Language Resources Association (ELRA), 2012.

[16] Hinrich Schütze. Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 251–258. ACL, 1993.